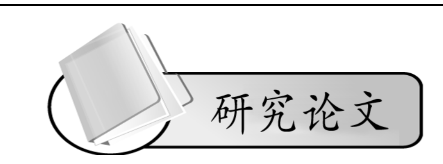


基于 IIPC 开源软件拓展构建国际重要科研机构 Web 存档系统

吴振新　张智雄　谢　靖　胡吉颖
(中国科学院文献情报中心　北京 100190)

摘要: 【**目的**】构建国际重要科研机构 Web 存档系统。【**方法**】基于 IIPC 开源软件拓展采集存档框架, 在采集端采用三层扩展策略, 在采集客户端增加自动上传及报告等管理功能, 开发 WARC 文件内容解析模块, 利用 Solr 进行索引。【**结果**】在采集端实现三层扩展, 通过增加采集客户端功能提高存档流程自动化程度, 通过增加的 WARC 文件内容解析功能抽取更多信息, 实现索引及检索服务的扩展。【**局限**】没有使用大规模采集存档进行检验。【**结论**】扩展后的采集存档框架初步具备分布式、可扩展、全自动化的特点。

关键词: 开源软件　网络信息存档　系统建设
分类号: G352

1　引　言

互联网资源被视为文化遗产的一部分, 受到许多国家立法认可。网络存档是对 Web 上的信息资源进行收集、保存并确保这些资源能够被长期使用的一系列持续活动, 为持久、有效地保存互联网资源提供了可靠的途径。截至目前, 全球近百个项目进行了 Web Archive 的研究和实践。

在科技领域, 大量的科技信息资源被发布在网络上, 近几年国际网络存档的焦点已经逐渐转移到对重要科技网络专题信息及科技机构网站的保存。2012 年 11 月美国国家数字信息基础设施保存计划(NDIIPP)发布《处于危险中的科学: 构建在线科学内容保存的国家战略》[1]报告, 明确将在线科学内容保存提升成为美国国家战略。

在这样的时代背景下, 重要网络科技信息资源已经成为科技信息资源建设体系中一种非常重要的开放资源, 这些资源的保存也成为科技战略资源保障工作中的一个重要组成部分。中国科学院文献情报中心作为国家级的保存机构, 充分意识到网络信息保存的重要性, 早在 2006 年就开始关注网络存档, 获得了国家社会科学基金的支持, 积极开展网络信息存档的研究与实践, 并于 2013 年正式开展国际重要科研机构 Web 存档。本文主要介绍在存档实践中如何基于 IIPC 开源软件拓展构建国际重要科研机构 Web 存档系统(NSL-WebArchive)。

2　IIPC 基本采集存档框架及应用分析

建于 2003 年的国际互联网保存联盟(International Internet Preservation Consortium, IIPC)[2], 目前已拥有超过 40 个成员机构, 涵盖来自世界各地的主要图书馆、档案馆、大学、非营利组织以及商业服务供应商, 在世界范围极大地促进了各国合作和交流共享。

IIPC 资助开发的用于网站遴选、收割和保存的各种网络存档工具[3]已经在全球得到了广泛部署和使用, 国内也有文章[4]对此进行介绍。目前使用范围最广的是覆盖了网络资源采集基本流程的 4 个工具包和已经

通讯作者: 胡吉颖, ORCID: 0000-0003-1559-2849, E-mail: hujy@mail.las.ac.cn。

成为国际标准的网络存档格式 WARC，笔者将其归纳如图 1 所示：

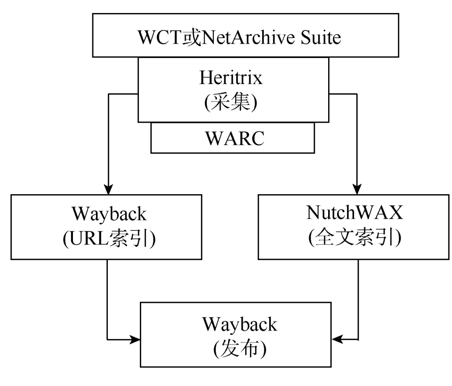


图 1　IIPC 基本采集存档框架

(1) WARC[5]：即 ISO 28500, Web 采集资源存档格式标准。

(2) Heritrix[6]：由 Internet Archive[7]牵头开发的、具有高度可扩展性的开源 Web 网络爬虫。

(3) Web Curator Tool(WCT)[8]：选择性网络采集过程控制及管理工具。

(4) Wayback[9]：提供基于 URL 的检索及访问的存档资源访问软件。

(5) NutchWAX[10]: Web Archive 全文索引工具。

国内由于 Web Archive 项目开展有限，鲜有利用这些开源工具进行大规模采集、存档、服务的案例，目前只有国家图书馆率先利用该框架[11]部署了实验系统并开展了多年的存档，为了解决运维的效率问题，在整个框架上增加了管理层，但依旧有许多环节、功能亟待扩展和完善，这些已经列在他们的发展规划中。国内还有一些针对 Heritrix 的研究，主要涉及扩展其核心的 5 个模块，分别实现抓取特定网站内容、调整采集策略提高抓取效率、对其进行改造实现增量式网络爬取等，多为研究性论文，缺少实用性系统的案例。

IIPC 联盟成员对其开源工具的应用非常广泛，也有多家机构开展了合作存档项目[12]，但对于分布式采集管理缺乏高效的开源管理软件，负责 WCT 开发的英国国家图书馆，自行研发了一套替代 WCT 的平台来管理 Heritrix 进行分布式采集，同时他们放弃了 NutchWAX，改用 Solr 对存档进行索引，法国国家图书馆的 Web Archive 项目也采用了类似的方案。

3　NSL-WebArchive 应用存档框架的个性化需求分析

虽然 IIPC 的采集框架在全球得到了广泛应用，但在实际的存档活动中，还需要结合个性化需求予以不同程度的个性化改造应用。

NSL-WebArchive 需要周期性地采集大量的科技网站资源，还要遵循网络礼仪，以较低的频率和速度进行采集，这样就存在大量资源的采集周期与采集速度之间的矛盾。同时大量的资源需要消耗大量的人力，那么自动化的需求随之提高。由于存档内容还需要进一步深度挖掘以提供分析服务，因此自然而然产生了大规模、分布式、自动化的采集及深度处理的个性化需求。在采用 IIPC 采集框架作为基础构建采集存档系统时，需要就这些个性化需求予以深入考虑，提出有效解决方案。

(1) 采集框架的平行扩展

NSL-WebArchive 的采集目标为相对固定和明确的网站群，需要对目标网站进行全域采集，因此可以使用轻量级爬虫提高任务运行效率并减轻采集服务器和被检测站点服务器的运行负担。

由于需要科技机构网站相对数量较多，还要遵循网络礼仪，以较低的频率和速度进行采集，因此，要在指定时间内完成大量采集任务，就需要部署大量的采集节点实施低频低速的分布式采集。同时，采集节点数量还应该能够根据任务需要进行扩充、收缩和动态调配。这就需要一个易于管理的采集端扩展策略，同时还需考虑在低频低速的采集模式下充分发挥服务器硬件的使用效率。

(2) 高效的分布式采集存档管理方案

NSL-WebArchive 分布式采集框架需要部署多个 Heritrix 采集实例以低频低速的采集模式完成大量采集任务，高效的分布式管理系统是采集存档平台必不可少的部分。

作为采集管理平台的 WCT，目前只能管理一个 Heritrix 实例，不能同时管理多个 Heritrix 实例。如果采集端不断扩展，即意味着部署多个 Heritrix 实例，而 WCT 却无法进行统一管理调度，这就引发了平行扩展后的采集管理问题，需要构建统一平台对分布式的采集节点实施采集以及存档任务的部署、管理。

另外如果部署多个 Heritrix 实例，每个实例的采集配置文档和产生的存档文件都需要修改缺省的命名规范以避免混淆，便于在统一管理时存档人员有效识别和管理这些文档。在考虑文档命名规则时要考虑采集文档来自不同的采集器，这些采集器部署在不同的服务器上，而且同一资源需要多次采集，这些信息都应予以有效记录。

(3) 高度的自动化流程

NSL-WebArchive 需要周期性地进行大量网站的采集，因此要求整个采集存档流程的自动化程度要大幅度提高。

①大量采集任务的配置、管理、周期性运行调度以及质量检验，这意味着需要大量的人工参与，需要实现采集任务管理的自动化。

②Heritrix采集的数据只能在本地指定的目录进行存储、管理，不能直接存放到远程存储目录，而分布式采集框架需要部署多个分布式的采集器，这就需要考虑平行部署多个采集实例后的资源收集问题。

③Wayback目前只能对指定的本地数据目录进行自动索引和提供浏览访问服务，无法同时为不同Heritrix实例采集的数据提供自动索引和浏览访问服务，即使在同一服务器上的多个Heritrix实例，由于各自存档目录不同，在人工归并之前，Wayback也无法为它们进行自动索引。

④目前NutchWAX需要将WARC放到Hadoop的文件系统中进行全文索引，不能进行本地索引，因此需要将不同Heritrix实例采集的数据统一集中到Hadoop中进行索引，索引后再将索引文件移回到Wayback目录下才能使用。这也是流程自动化需要考虑的一个问题。

(4) 丰富的内容与服务方式

由于国际重要科技网站资源存档是面向学科的国际重要科技机构网络资源保存，资源保存之后更为重要的是为用户提供基于内容的深度挖掘和分析服务，因此系统不但要有基本的网站 URL 的检索和浏览功能，还要有多角度多层次的内容提供、内容分析服务的能力。

NSL-WebArchive 需要考虑存档内容信息的抽取，增加索引维度，提供包括学科、时间、站点在内的分面浏览和全文检索，解决目前存档内容索引不足和访问服务单一的问题。

Wayback 只提供基于 URL 的索引和检索，而按照用户的使用习惯和需求，这样单一的功能是远远不够的。如上文所述，目前提供全文索引的 NutchWAX 需要将 WARC 文件和索引文件在 Heritrix、Hadoop、Wayback 之间往返移动，同时实践中发现它在性能上存在一定的问题，对硬件有较高的要求。

4 NSL-WebArchive 平台的关键问题解决方案

基于上述个性化需求分析，笔者提出基于 IIPC 采集存档框架的构建思路：

(1) 提供一个面向大规模采集的可扩展的系统框架，从多层面实现系统的可扩展性。

(2) 将相关工具作为框架中的组件纳入，不改变开源工具本身的功能。

(3) 通过中心管理端和客户采集端的模式，实现分布式采集存储，支持多节点协同工作，并充分利用硬件与网络完成采集任务。

(4) 完善客户端管理软件功能，提高采集流程的自动化程度。

这样既可快速实现采集系统平台，还能够充分利用原有工具的优点，同时具有更好的兼容性，可实现无缝升级，尽享开源工具的优势。

4.1 分布式 Web 存档框架整体方案

NSL-WebArchive 设计了中心管理端和客户采集端的模式以实现分布式框架，如图 2 所示。框架主要由三种类型节点组成：管理节点、采集节点和存储(索引访问)节点。

(1) 管理节点

采集管理平台负责完成种子站点采集配置和管理，同时还负责采集任务生成和采集任务队列的管理，提供查询各采集任务的执行情况。

(2) 采集节点

每一个采集节点都部署了一个客户端，一个客户端管理一个对应的 Heritrix。根据 Heritrix 的采集任务完成情况到采集管理平台的任务队列中获取采集任务，从管理数据库中取得相关信息生成 Heritrix 所需的相关文档，并依据任务要求调度 Heritrix 对 Internet 上指定站点进行采集，将采集结果以 WARC 格式进行存储。在采集任务结束后，以 FTP 方式将完整的 WARC 文件传输到存储节点的指定目录中，将任务完成情况存入管理数据库供采集管理平台查询使用。

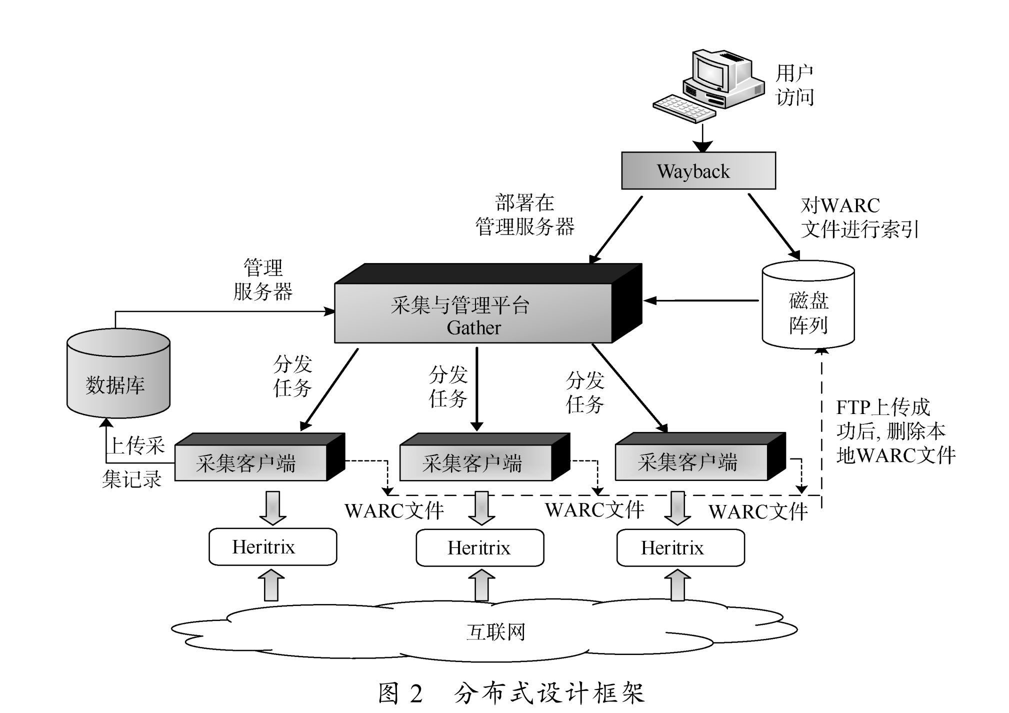


图 2　分布式设计框架

(3) 存储(索引访问)节点

存储节点除了存储从采集节点推送过来的 WARC 文件，还要部署 Wayback 和存档内容抽取模块以及索引辅助工具，提供对存档资源的索引和访问的功能。

4.2　基于 Heritrix 的分布式采集扩展框架

NSL-WebArchive 需要一个面向大规模采集的分布式、自动化的采集方案，为了实现这个目标，笔者提出一个三层扩展策略以提高 NSL-WebArchive 采集系统的可伸缩性，如图 3 所示：

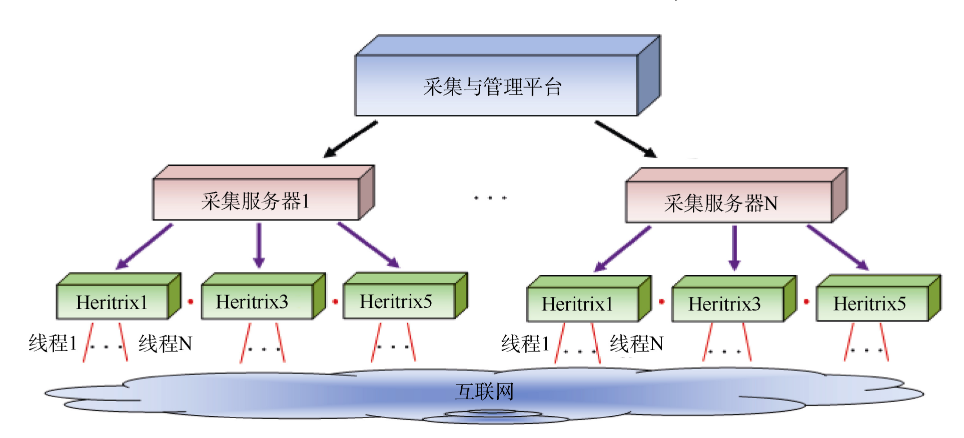


图 3　基于 Heritrix 的分布式采集扩展框架

(1) 采集服务器的水平扩展。即最简单的方法，增加采集服务器。

(2) 采集节点的水平扩展。在不影响采集效率的情况下，同一采集服务器上部署多个采集节点，即部署多个 Heritrix。

(3) 采集线程的水平扩展。利用 Heritrix 多线程的特点，改变队列算法，同时启动多个线程采集多个站点。

原理上，三层扩展策略是简单的平行扩展策略，实施层面，则需要综合考量多种因素才能确定具体指标，如平衡每个采集器采集的种子站点体量、次均采集速度、完成时间，并多次测试观察每台服务器硬件使用效率，确认合适的采集器部署数量和实施采集的线程数。

4.3　分布式 Web 存档系统的基本流程

分布式 Web 存档系统的基本流程如图 4 所示。

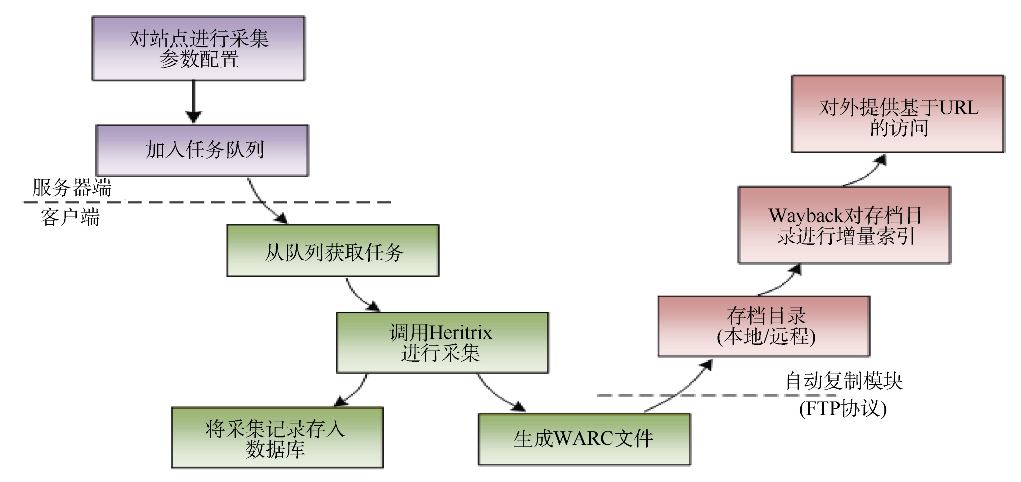


图 4　分布式内容采集系统的基本流程

存档人员在采集管理平台上配置管理种子站点，管理平台按照配置自动将采集任务适时加入任务队列。

采集节点的客户端程序监控 Heritrix 的状态，主动到采集管理平台的任务队列中接收任务。调用 Heritrix 进行网页内容的采集，采集结果以 WARC 格式存储。当每一个采集任务结束后，采集客户端自动将生成的 WARC 文件通过 FTP 传送到指定的存储节点目录下按年月进行分类存储。当上传成功后，采集客户端删除本地的 WARC 文件。Heritrix 对每次任务都会生成一系列报表，记录此次网页采集情况，当采集任务结束后，采集客户端从 Heritrix 的报表中提取出此次采集完成情况的一些关键参数存入管理数据库，供管理端查看站点采集的历史记录。

部署在存储节点的 Wayback 会自动对指定目录进行检查，对监测到的新存入的 WARC 文件进行 URL 索引，索引完成后用户就可以通过 Wayback 对存档资源进行基于时间轴的访问。

4.4　采集节点主动模式

采集框架通过部署一个中心管理服务器和多个客户端采集服务器实现分布式采集，其最大的亮点则是采集节点的主动工作模式。该模式原理如图 5 所示。部署在采集节点的客户端采用 RMI[10]远程调用方式，在采集管理平台与采集节点之间建立安全稳定的通信管道，主动获取采集任务并上报采集状态，使得采集管理平台不需要轮询众多的采集节点，可以有效地减少采集机器故障对整体采集系统造成的影响。当一台采集节点计算机出现故障时，该节点将停止向服务器申请新的采集任务，该节点的采集任务均匀分派到其他的采集节点上。虽然会降低采集效率，但不会造成对目标站点采集失败的情况出现，可以保障数据采集的时效性和准确性。

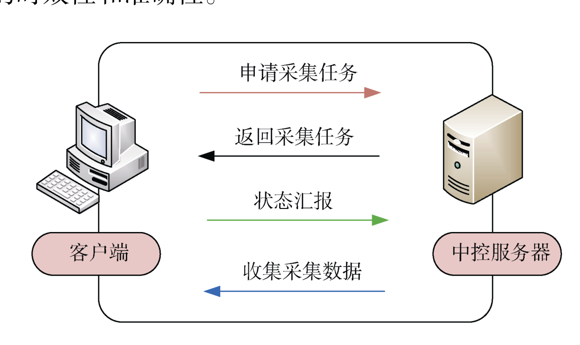


图 5　采集节点与采集管理平台的通信模式

采集信令是中心管理服务器向采集节点发送信息的标准指令，包含完整的采集任务描述。在一次通信中，采集节点从中心管理服务器领取最新采集任务，根据信令内容调度、控制 Heritrix 完成采集任务。采集信令是分布式采集框架的关键设计环节，其中包括：

(1) 站点的唯一标识ID：用于中控服务器统一分配回收采集任务。

(2) 采集入口 URL (即 Seed URL)：通知采集节点从此 URL 开始采集任务。

(3) 采集限定范围：告知采集节点采集网站子域名、子目录的规则以及域名外链 URL 的采集规则，采集节点明确规则以外的 URL 停止采集。

(4) 采集速度及服务器压力参数：采集一个站点的线程数，采集一个 URL 的延迟时间(以缓解对方服

务器压力)。

(5) 站点采集频率：根据站点的内容发布周期确定多长时间对站点完成一次采集，归档存储站点新发布或者修改的内容页面。

(6) 针对目标站点的个性化配置：包括连接响应等待时间、回话保持时间、最大下载文件限制、Cookie管理等，以目标站点规定的合法方式获取数据。

4.5 功能扩展以提升流程自动化程度

(1) 中心管理服务器的自动任务调度

NSL-WebArchive需要进行大量的采集任务配置、管理、周期性运行调度以及质量检验，如图6所示，中心管理平台通过任务调度机制实现了大量任务周期性运行调度的自动化，基本上种子站点经过一次配置即可开始周期性采集。同时能够对采集线程实时监控状态，及时收集和分析采集日志。

图6 中心管理平台

这种任务调度机制需要管理员指定每个站点的采集设置，包括采集深度、采集频率、最长的采集时间、最大的下载量、最大跳转次数、路径最大深度等。管理服务器通过设置定时器实现周期性的采集任务生成。

(2) WARC文档自动汇集策略

采集客户端采用模块化的工作流机制循环执行采集任务，从主动获取采集指令到发送采集结果报告，整个流程包括多个功能模块。客户端增加了自动收集归并模块解决WARC文件的自动收集问题。

采集客户端监测到采集任务结束后，即调用归并模块将WARC文件通过FTP方式传送到指定的存储节点目录下，上传成功后，将本地的WARC文件删除，同时还将Heritrix生成的job文件夹下相应的日志文件也通过FTP方式传送到管理服务器的job文件存储目录下。

该模块的增加不但解决了Heritrix远程存储的问题，也解决了利用多个Heritrix进行分布式采集时的结果汇集问题。

(3) WARC文档分散存储策略

长期采集大量站点，必须对采集资源实施存储管理，因此笔者制定了存档服务器的存档目录按年月进行分目录存储，确保每个存档目录不会太大，便于长时间的保存及备份管理。

(4) 错误报告机制

采集客户端增加了采集情况报告模块，在采集任务完成或中断后(如采集陷阱、种子变更导致中断)，用于自动分析Heritrix的采集日志，包括：

①基本采集情况报告模块：当采集任务结束后，读取采集任务文件夹下Heritrix生成的crawl-report，获取采集所用时间、采集成功URL数量、采集失败URL数量、下载数据量等信息，并将这些信息存入管理数据库。

②高级采集情况报告模块：为了查看更加详细的采集情况，如HTTP状态码对应的URL数量和所占比例、采集的文档类型对应的URL数量和所占比例、种子采集情况、采集URL列表和一些错误信息等，本文对Heritrix采集报告部分的源代码做了部分修改，利用每次生成任务的job ID号构造URL链接到Heritrix的统计界面，查看每次任务采集的详细信息统计情况。

4.6 构建规范的文件命名体系

分布式Web存档系统中有4种文档命名需要规范。

(1) 种子文件命名

对每个站点进行采集都会生成一个种子文件，用来保存采集的站点地址，Heritrix根据种子文件确定要采集的站点。

种子文件的命名方式定为：站点域名-seeds.txt。

(2) 配置文件命名

对每个站点进行采集参数的配置都会生成一个配置文件，Heritrix根据配置文件对站点进行采集。

配置文件的命名方式定为：站点域名.xml。

(3) 任务文件夹与任务文件命名

每个任务都会生成一个任务文件夹，存放采集的日志信息和报表等。为了更好地管理任务，需要对任务实现按月存放，因此在任务文件夹下按时间年月生成新文件夹用来存放当月的采集任务。

每个文件夹的命名方式定为：年月。如：201403、201404、201405。201403文件夹下存放2014年3月采集任务生成的任务文件夹。

任务文件的命名方式定为：站点域名-时间戳。

需要说明的是，每个站点域名是设置的任务名称；

任务文件夹生成时间采用 UTC 时区(加 8 个小时是北京时间), 格式为: yyyyMMddHHmmss。

(4) WARC 存档目录及 WARC 文件命名

在存储节点上, WARC 文件被指定存放在/mnt1/WARCs 下。为了实现按月存放, 需要采集客户端在上传数据时, 在/mnt1/WARCs 下自动新建文件夹, 每个文件夹以存档的年月命名, 如: 201403、201404、201405。201403 文件夹下存放 2014 年 3 月采集的所有 WARC 文件。

WARC 文件的命名方式定为: 站点域名-WARC 文件生成时间-序列号-采集机器的 Hostname。

①每个站点的域名是 WARC 文件的前缀;

②WARC 文件生成时间采用 UTC 时区, 加 8 个小时是北京时间, 格式为 yyyyMMddHHmmss;

③序列号是每一次采集任务生成的多个 WARC 文件的序号。由于预先定义了 WARC 文件的大小, 如限定一个 WARC 文件大小不能超过 1 GB, 当一次任务采集的数据小于 1 GB 时就只有一个 WARC 文件, 序号为 0000; 当大于 1 GB 将被拆分为多个文件, 顺序采用 0001、0002, 以此类推;

④采集机器的 Hostname 是 WARC 文件的后缀。

例如, 要采集的站点的英文名称为 International Association for Hydrogen Energy, 网址是 http://www.iahe.org/, 按照上述命名方式生成各类文件命名如下:

(1) 种子文件: www.iahe.org-seeds.txt;

(2) 配置文件: www.iahe.org.xml;

(3) 任务文件夹: www.iahe.org-20140323084011;

(4) WARC 文件: www.iahe.org-20140323084024-0000-Hadoop-master-180.WARC.gz。

4.7　基于 WARC 内容抽取的内容获取服务扩展

考虑到检索服务和浏览服务的扩展需求, NSL-WebArchive 利用 Wayback 底层代码中对 WARC 文件内容进行解析的三个类: WARCReaderFactory、WARCReader、WARCRecord, 将其中的 get(String warcFilePath)、getHeader()、read()方法分离封装成独立的模块, 用于 WARC 文档内容解析与抽取, 具体技术实现细节将另外撰文详述。

目前 NSL-WebArchive 已经初步实现对 WARC 文件内容的抽取, 利用 Solr 技术建立基于内容的分面索引, 实现对存档资源基于内容的检索, 同时利用中心管理平台的种子描述内容, 还能提供基于时间、学科、资源类型的存档站点分面功能, 如图 7–图 9 所示。利用管理统计信息, 为每个种子站点提供采集统计信息。基于 WARC 的内容抽取丰富了检索和访问服务, 也为今后提供基于内容的深度挖掘和分析服务打下良好基础。



图 7　NSL-WebArchive 的访问服务首页

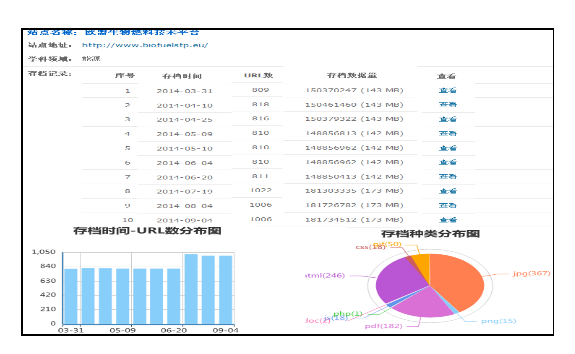


图 8　NSL-WebArchive 的站点浏览页面



图 9　NSL-WebArchive 的分面检索结果浏览页面

5　NSL-WebArchive 平台运行效果分析

目前整个 NSL-WebArchive 平台基本完成, 228 个科技网站进入周期采集存档。截至 2014 年 9 月, 存档数据总量为 1.1 TB (压缩), WARC 文档总数为 1 200, 存档 URL 总数为 11 392 701, 采集资源格式分布如图 10 所示。根据 2014 年 9 月的采集日志, 有 170 个站点可以自然结束, 有 58 个站点被规则中断(爬行时间>7 天或下载量>70GB), 需要人工查看判断。在 24 个采集

客户端同步采集的情况下，约 40 天左右时间完成一次采集。

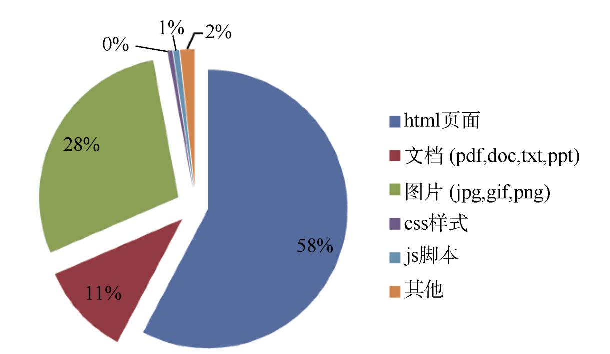


图 10　总体存档资源格式分布图

以 2014 年 10 月 27 日采集的 www.iiasa.ac.at 为例:

(1) 试图采集 URL 总数: 49 507。

(2) 采集成功URL数: 48 517，主要状态如表1所示:

表 1　采集成功的 URL 主要状态分析

状态码	URL 数量	比例	状态码含义
200	43 010	88.8%	请求已成功，请求所希望的响应头或数据体将随此响应返回
404	2 466	5.1%	没有找到，请求失败，请求所希望得到的资源未在服务器上发现
301	1 508	3.1%	跳转，被请求的资源已永久移动到新位置
302	1 301	2.7%	跳转，请求的资源现在临时从不同的 URI 响应请求
403	108	0.2%	禁止，服务器已经理解请求，但是拒绝执行

(3) 采集失败URL数: 990，原因有待进一步分析。

(4) 采集所用时间: 3d23h12m53s937ms (约 4 天)。

(5) 采集数据量: 20 GB。

(6) 平均每秒下载量: 62 KB。

(7) 爬行 Host 数: 24。

(8) 采集格式分析如图 11 所示:

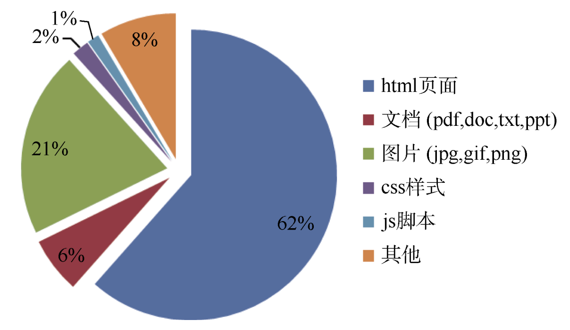


图 11　示例网站一次采集资源格式分布图

总体上取得了较好的效果:

(1) 采集管理系统提供了丰富的站点管理信息，基本上做到一次配置即可周期性采集，可以对种子站点的采集效率实现稳定控制，实现了采集存档的有效管理。

(2) 扩展框架从理论上看，采集的三个层面可以实现无限平行扩展。目前在 4 台服务器上部署了 24 个采集节点，实现了分布式采集，并保障了良好的可扩展性。

(3) 通过客户端功能的扩展，NSL-WebArchive 大幅度提高自动化程度，在采集任务配置、周期性运行调度、WARC 文档汇集等方面极大地减少了人工参与，基本上实现了包括任务部署、采集、存档、索引、访问这样一个完整的自动化流程。

(4) 封装了 WARC 文档解析模块，初步实现对 WARC 文件内容的全面抽取，可以建立基于内容的分面索引，不但丰富了检索和访问服务，也为今后提供基于内容的深度挖掘和分析服务打下良好基础。

目前 NSL-WebArchive 平台的种子站点数量逐步增加，还需要通过采集存档更大规模的数据以检验整个框架的稳定性、效率以及扩展性。

6　结　语

通过持续的网络存档，不但实现(科技)文化遗产的完整保存，同时还可以对存档资源进行深入分析、挖掘和再利用，支持相应科技政策和技术的效果评估、长期科技战略决策、领域变化趋势分析、预测未来发展趋势等，从而利用存档的网络资源更好地为学术研究和社会发展服务。

通过国际重要科技机构网络信息存档系统的建设，可以为科技网络信息资源初步提供可靠的保存体系，对长期地利用存档资源为学术研究、情报人员、科技管理人员提供服务提供有利的支撑。通过 NSL-WebArchive 平台建设，在大规模网络存档的可管理、易扩展、自动化和信息抽取挖掘等方面进行了初步探索。目前系统在不断扩大存档规模的同时，也开展存档信息内容的深度挖掘、分析和再利用研究，考虑开展多种基于 Web 数据分析的情报支持服务，同时也在考虑如何检验和评价 Web 存档的完整性和真实性，以确保存档资源的可信赖。

参考文献：

[1] Toward a National Strategy for Preserving Online Science [EB/OL]. [2014-08-05]. http://www.digitalpreservation.gov/meetings/documents/othermeetings/science-at-risk-NDIIPP-report-nov-2012.pdf.

[2] IIPC [EB/OL]. [2014-08-05]. http://netpreserve.org/.

[3] Tools and Software [EB/OL]. [2014-08-05]. http://netpreserve.org/Web-archiving/tools-and-software.

[4] 刘兰, 吴振新, 向菁, 等. 网络信息资源保存开源软件综述[J]. 现代图书情报技术, 2009(5): 11-17. (Liu Lan, Wu Zhenxin, Xiang Jing, et al. Review of Open Source Software in Web Archive [J]. New Technology of Library and Information Service, 2009(5): 11-17.)

[5] ISO 28500:2009 Information and Documentation -- WARC File Format [EB/OL]. [2014-08-05]. http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=44717.

[6] Heritrix [EB/OL]. [2014-08-05]. https://Webarchive.jira.com/wiki/display/Heritrix/Heritrix.

[7] Internet Archive [EB/OL]. [2014-08-05]. http://www.internetarchive.org/.

[8] The Web Curator Tool Project [EB/OL]. [2014-08-05]. http://Webcurator.sourceforge.net/.

[9] Web Archive Access [EB/OL]. [2014-08-05]. http://sourceforge.net/projects/archive-access/files/wayback/.

[10] NutchWAX [EB/OL]. [2014-08-05]. http://archive-access.sourceforge.net/projects/nutch/.

[11] 吴振新, 曲云鹏, 李成文, 等. 基于开源软件搭建网络信息资源采集与保存平台[J]. 现代图书情报技术, 2009(7-8): 6-10. (Wu Zhenxin, Qu Yunpeng, Li Chengwen, et al. Constructing a System for Harvesting and Preserving Chinese Web Information Resources Based on Open Source Software [J]. New Technology of Library and Information Service, 2009(7-8): 6-10.)

[12] Trail: RMI [EB/OL]. [2014-08-05]. http://download.oracle.com/javase/tutorial/rmi/index.html.

[13] 吴振新,张智雄,王婷.网络信息资源保存的协作网络研究[J]. 数字图书馆论坛. 2009(7): 2-6. (Wu Zhenxin, Zhang Zhixiong, Wang Ting. Research on the Web Archive Cooperative Networks [J]. Digital Library Forum, 2009(7): 2-6.)

作者贡献声明：

吴振新：系统框架设计及实施管理，论文撰写；
张智雄：提出扩展思路，完善系统框架设计；
谢靖，胡吉颖：系统开发，论文撰写。

收稿日期: 2014-09-03
收修改稿日期: 2014-11-03

Developing Web Archive System of International Institutions Based on IIPC Open Source Software

Wu Zhenxin Zhang Zhixiong Xie Jing Hu Jiying
(National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: [**Objective**] Develope Web Archive System of International Institutions. [**Methods**] Based on IIPC open source software framework, this paper applies a three layer expansion strategy in the acquisition terminal, provides automatical uploading and reporting function in the acquisition client, develops a WARC parser which can analyze the content of WARC file, uses Solr to be an indexer. [**Results**] This paper implements acquisition expansion, promotes the automatical level of system workflow by adding more function modules in the acquisition client, extracts more information by developing WARC parser modules, uses Solr to enrich index and retrieval service. [**Limitations**] Lack of large-scale Web archive to verify this platform. [**Conclusions**] The expanded Web archive framework becomes distributed, extended and full automatic.

Keywords: Open source software Web archive Syetem development